PB86-213204

Non-Neoplastic Lesions: Use of Data from Pre- or Non-Neoplastic Lesions That May Indicate Potential for Carcinogenesis

(U.S.) Health Effects Research Lab.
Research Triangle Park, NC

May 86

U.S. Department of Commerce
National Technical Information Service
NTIS

PB86-213204

EPA/600/D-86/114
May 1986

NON-NEOPLASTIC LESIONS: USE OF DATA FROM PRE- OR NON-NEOPLASTIC LESIONS THAT MAY INDICATE POTENTIAL FOR CARCINOGENESIS

by

Judy A. Stober
Toxicology and Microbiology Division
Health Effects Research Laboratory
U.S. Environmental Protection Agency
Cincinnati, Ohio 45268

HEALTH EFFECTS RESEARCH LABORATORY
OFFICE OF RESEARCH AND DEVELOPMENT
U.S. ENVIRONMENTAL PROTECTION AGENCY
RESEARCH TRIANGLE PARK, NORTH CAROLINA 27711

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA/600/D-86/114		PB86-213204/AS

4. TITLE AND SUBTITLE	5. REPORT DATE
Non-Neoplastic Lesions: Use of Data from Pre- or Non-neoplastic Lesions That May Indicate Potential for Carcinogenesis	May 1986
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
Judy A. Stober	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELEMENT NO.
U.S. EPA HERL/TMD/CSSB 26 W. St. Clair St. Cincinnati, Ohio 45268	CBMC1A
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
U.S. EPA HERL/TMD/CSSB 26 W. St. Clair St. Cincinnati, Ohio 45268	Workshop
	14. SPONSORING AGENCY CODE EPA/600/11

15. SUPPLEMENTARY NOTES

P.O. Judy A. Stober CSSB/TMD/HERL/ORD/USEPA

16. ABSTRACT

The Toxicology and Microbiology Division of the US EPA, Health Effects Research Laboratory has initiated a research program to develop a matrix of short-term tests to distinguish carcinogens from non-carcinogens among genotoxic substances and to develop methods for predicting relative carcinogenic potency based on the results of short-term in vivo bioassays. This paper includes a basic description of the project and outlines the proposed statistical methods for evaluating the Carcinogenesis Testing Matrix (CTM). Successful development, testing and implementation of the test matrix system is very dependent on statistical design, analysis and prediction procedures. Since this is an extension of the usual application of a single screen to predict a qualitative effect, considerable statistical development and testing is needed. Three short-term in vivo bioassays have been proposed for the CTM, the sencar mouse skin bioassay, the mouse lung adenoma bioassay and the liver foci bioassay. Appropriate statistical methods in terms of statistical power and robustness for analyzing the data from each assay are being developed and evaluated. Methods for estimating and comparing potencies from the long-term studies and short-term test systems are being developed and evaluated. The proposed methods include procedures for comparing potencies, estimation of confidence limits and assessment of statistical significance. The methods will be evaluated in terms of statistical considerations, adaptability to laboratory test situations, ease of application, efficiency and predictive power.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
Release to Public	Unclassified	25
	20. SECURITY CLASS (This page) Unclassified	22. PRICE $9.95

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

NOTICE

This document has been reviewed in accordance with U.S. Environmental Protection Agency policy and approved for publication. Mention of trade names or commercial products does not constitute endorsement or recommendation for use.

NON-NEOPLASTIC LESIONS:
USE OF DATA FROM PRE- OR NON-NEOPLASTIC LESIONS THAT MAY INDICATE POTENTIAL FOR CARCINOGENESIS

Judy A. Stober
U.S. Environmental Protection Agency
Health Effects Research Laboratory
Cincinnati, Ohio

Under various congressional acts and statutes the U.S. Environmental Protection Agency is required to set standards for harmful pollutants in water. To meet the needs of the regulatory office and support regulations, extensive and reliable toxicological data bases must be developed and maintained. Traditionally, in the absence of appropriate human epidemiologic data, estimates of carcinogenic risk to humans are based on lifetime carcinogenesis bioassays. The applicability of long-term bioassays to environmental samples is limited by the time and expense required for the assays and the unknown and variable composition of the environmental samples. Thus, there is a need to develop more efficient and economical tests that can reliably predict a chemical's potential to cause chronic effects.

The Toxicology and Microbiology Division of the U.S. EPA Health Effects Research Laboratory has initiated a research program to develop a matrix of short-term tests to distinguish carcinogens from non-carcinogens among genotoxic substances and to develop methods for predicting relative carcinogenic potency based on the results of short-term in vivo bioassays. Successful development, testing and implementation of the test matrix system is very dependent on statistical design, analysis and prediction procedures. Since this is an extension of the usual application of a single screen to predict a qualitative effect, considerable statistical development and testing is needed. Appropriate statistical methods for analyzing the response data from each short-term assay for a positive dose related response are being developed and evaluated in terms of power and robustness. Statistical methods to estimate carcinogenic potency from the short-term assays and long-term assays have been proposed. The proposed methods include procedures for comparing relative potencies, estimation of confidence limits and assessment of statistical

significance. The methods will be evaluated in terms of statistical considerations, adaptability to laboratory test situations, ease of application, efficiency and predictive power.

A brief description of the Carcinogenesis Testing Matrix project (CTM) will be presented. The CTM project was designed to be completed in stages with each subsequent development based on the results and thorough evaluation of the previous research. The focus of this paper is the proposed statistical methods to be used in the evaluation of the the initial phase of the work.

CARCINOGENESIS TESTING MATRIX

The Carcinogenesis Testing Matrix research project was originally presented and reviewed in a workshop sponsored by the Health Effects Research Laboratory of the U.S. Environmental Protection Agency. At the "Workshop on Short-term in vivo Assays for Environmental Carcinogens", Bull and Pereira discussed the objectives, design and, in general, the experimental approach for the validation of the CTM (1). A three tiered testing system was proposed. Tier 1 was designed to be a screening procedure to select with a minimum of false negatives those chemicals which have high potential for being mammalian mutagens and carcinogens. The Ames Salmonella mutation assay and various mammalian cytogenetic assays were designated as Tier 1 tests. Tier 2 was defined as the Carcinogenesis Testing Matrix, a collection of short-term tests that measure response in terms of benign lesions. Tier 2 would be used to "confirm" carcinogenic activity indicated in Tier 1, i.e., eliminate false positives from Tier 1 with a minimum of false negatives, and to predict relative carcinogenic potency of individual chemicals or complex mixtures. Tier 3 was defined as the life-time bioassays and designated for situations for which there existed a need and a capability of quantitative risk estimation.

Criteria were defined for the short-term bioassays to be considered for inclusion in the CTM. The focus was on a battery of tests in combination as a matrix system for prediction, not on individual tests; each test being potentially selective in its response to certain chemicals or chemical classes. Of primary importance was the requirement that the response measures of the short-term assays have to be clearly related to oncogenesis and that the short-term tests have to be capable of detecting the majority of carcinogens. Three bioassay systems were selected for the first round of testing and their potential as Tier 2 test systems was discussed in numerous papers presented at the "Workshop on Short-term in vivo Assays

for Environmental Carcinogens" (2). Three assays, the mouse skin initiation/promotion bioassay, the rat liver altered enzyme foci bioassay, and the mouse lung adenoma bioassay are currently being evaluated as a battery of short-term Tier 2 tests. Eighteen known carcinogens have been selected for testing in the first round evaluation of the CTM (Table 1). They were chosen for their diversity of chemical structure and target organ specificity. Thus, the emphasis of the initial testing is on evaluating the false negative rate of the proposed short-term tests and their ability to predict relative carcinogenic potency for a select group of compounds. At this stage no information will be available on the false positive rate of the short-term assays.

Table 1

INITIAL CTM COMPOUNDS

Aflatoxin B_1	Epichlorohydrin
2-acetylaminofluorene (2-AAF)	N-(4-(5-Nitro-2-furly)-2-thiazolyl)-formanide (FANFT)
Azaserine	3-methylcholanthrene (3-MC)
Azobenzene	Methylmethanesulfonate (MMS)
Benzene	2-naphthylamine
Benzo(a)pyrene (BAP)	β-propiolactone
1,2-dibromoethane (1,2 DBEANE)	Safrole
Dibutylnitrosamine (DBNA)	Urethane
Dimethylhydrazine (DMH)	
Diethynitrosamine (DENA)	

STATISTICAL METHODS FOR VALIDATION OF THE CTM

There are four areas in the validation of the Carcinogenesis Testing Matrix project which require development and evaluation of statistical methods:

1. development of robust sensitive statistical methods to evaluate each of the short-term assays for a positive dose related response,
2. development of statistical methods to estimate carcinogenic potency from long-term bioassay,

3. development of statistical methods to estimate carcinogenic potency from the short-term assays, and
4. methodology to evaluate the ability of the CTM to reliably predict relative potency.

The proposed statistical methodology to be used in each area will be presented and briefly discussed in the following sections.

Statistical Methods for Short-term Bioassays

Although short-term bioassays are widely used to screen for toxicity, relatively little work has been done to determine the best methods to test statistically for positive dose related responses in the assays. The objective for the CTM project is to define appropriate, sensitive and useful statistical techniques for analyzing data obtained from the Sencar mouse skin initiation/promotion assay, the Strain A mouse lung adenoma assay and the rat liver altered enzyme foci assay. The primary emphasis of the short-term assay is the testing for chemical carcinogens and tumor initiators and promoters. The secondary objective is the estimation of a dose response model. The development of the statistical methods includes a review of each of the proposed screening assays, review of available statistical methods reported in the literature and development of alternative statistical procedures as required. The proposed methods are being evaluated using available data generated by staff scientists conducting the short-term assays on the initial CTM compounds and simulation studies using data from representative studies as well as data generated to test the methodology under specific conditions.

Most short-term assays have a definable and easily quantifiable single response endpoint that is representative of specific biochemical phenomena. For each of the CTM short-term test systems, the aim was to determine intuitively reasonable statistical methods that would investigate the experimental hypothesis in a straightforward manner using all relevant information available. Unfortunately, the characteristics of the response data do not necessarily lend themselves readily to the classical techniques for statistical analysis. If an objective of Tier 2 testing is to verify the carcinogenic activity indicated in Tier 1, then the recommended methods should have sufficient power to detect biologically meaningful differences while minimizing the probability of a false positive. Various response measures and statistical tests, both omnibus and directional, have been proposed for the three assays. Each has been or will be evaluated in terms of mathematical statistical considerations, efficiency, robustness, power and ease

of application. The results and recommendations to date will be briefly summarized in the following sections.

Sencar Mouse Skin Painting Assay. Chemical carcinogenesis is currently viewed as a multistep process and the initiation/promotion system using the Sencar mouse skin model has been shown to be a predictive bioassay for certain classes of carcinogens. A general description of the Sencar skin bioassay can be found in Slaga, et al. (3) and Bull, et al. (4). Female Sencar mice between six and eight weeks of age are randomly assigned to treatment groups of sizes twenty-five, thirty or forty mice. In the initiation/promotion study, the tumor initiating potential of a substance is tested by exposing the animals by either single or multiple dosing of the test sample. The dosage is applied either topically or subcutaneously to the dorsum, orally or intraperitoneally. Two weeks following the final initiating dose a tumor promoter regimen is started and maintained for a minimum of twenty weeks. A control group receives an equivalent amount of the vehicle that was used for the administration of the initiator. The backs of the animals are shaved three days prior to the administration of the test sample and once weekly during the tumor promotion period to facilitate the administration of the test compounds and tumor observation.

At the Toxicology and Micrbiology Division of the Health Effects Research Laboratory a tracking system has been developed and implemented for the collection of data in the mouse skin assay system. In addition to the daily observation of the animals for health monitoring and morbidity, a weekly observation of all animals on study is made and all grossly observed tumors are individually charted as to location, size and type (papilloma or carcinoma). Only those tumors that are at least one millimeter in diameter and observed for a minimum of three consecutive weeks are included in the tumor count. All tumors are charted until they regress, coalesce or the animal is sacrificed.

The susceptibility of the Sencar mouse to the development of skin tumors and the possibility of tumor regression make the pattern of tumor occurrence an important consideration. Traditionally, mouse skin painting data have been analyzed using the proportion of animals with tumors in a treatment group or the total number of current tumors per animal at a specfic study time period as the response measure. These measures ignore survival differences in the groups and the pattern of tumor development, each of which may contain important information on the toxicity of a compound. The result may be a loss of statistical power, i.e. test sensitivity. In developing the methodology to be used in the analysis of tumor data from the

tracking system, the multiplicity of tumors as well as their time of occurrence were used as indicators of carcinogenicity. The objective was to define summary response measures and statistical methods that would take into account simultaneously survival rates and the total tumor occurrence pattern.

Table 2

Tumor Response Measures

Measure	Comments
Univariate:	
Time to 1st Tumor	Ignores Multiplicity of Tumors
Total Number of Current Tumors	Ignores Time of Tumor Occurrence, Duration and Tumor Regression
Total Number of "Cumulative" Tumors	Ignores Time of Tumor Occurrence and Duration
Integral w.r.t. time of the Number of Tumors	Takes Duration Into Account But Not Exact Time of Occurence
Weighted Sum of Tumors	Each Tumor Weighted by Its Time of Appearance. Ignores Tumor Regression
Multivariate:	
Multiple Times to Tumor	Takes Multiple Tumors Per Animal Into Consideration

Numerous approaches to testing for a dose-related effect were investigated by Crump and Ng (5). One class of tests was obtained by defining for each animal a single quantitative characteristic which reflects the tumor response in the animal and applying a univariate test to the summary response characteristic. The response measures investigated by Crump and Ng are listed in Table 2.

The univariate response measures are calculated for each animal over the entire study time of interest with their value censored at the time of the animals death. The multivariate response measure models multiple times to tumor without first reducing the response to a single summary value. It is recognized that the univariate response measures have inherent weaknesses. The time to first tumor ignores the multiplicity of tumors; the total number of current or cumulative tumors do not take the time of occurrence or duration into account; the integral weights each tumor by its duration but ignores the specific time of occurrence; the weighted sum of tumors weights each tumor by its time of occurrence but ignores tumor regression. Nevertheless, the univariate response measures are advantageous for several reasons: the analysis includes every animal not just animals surviving to a given time; tumor multiplicity is accounted for in all but one of the measures; and the duration and regression of tumors by several of the measures.

Both multiple comparison and trend type procedures were evaluated to test the various response measures for a dose-related effect. The data analysis is complicated by the occurrence of early deaths in the experimental animals. If it is assumed that each univariate response measure is censored at the value attained at the time of death, these data can be analyzed using methods suitable for survival data. Multiple comparison statistical tests which are appropriate for such data include the Gehan and Wilcoxon-type multiple comparison test (6,7,8) and the logrank test (8,9,10), a Cox-type multiple comparison test. These multiple comparison procedures are omnibus test in the sense that they have power to detect nearly all types of differences. A test for a dose-response trend can be made using a one-sided directional test based upon Cox's proportional hazard model (11). The multiple times-to-tumor data can be analyzed using a regression procedure based on a generalization to multivariate data of Cox's proportional hazard model and a partial likelihood (12).

The performance of the statistical procedures discussed above were examined using data collected on the skin painting biossay tracking system and data generated in a simulation study (5,13). For the study data, the Gehan Wilcoxon multiple comparison test had greater power than the comparable Cox-type test when applied to the

univariate summary response mesaures. In some cases, in the analysis of the time to first tumor data the Cox type trend test achieved a higher level of significance than either of the multiple comparison tests. The Gehan Wilcoxon test is designed to cover a wide range of alternatives and thus, could be more significant than a trend test in the absence of a monotone dose response. The comparison of the univariate response measures found little difference in the power of the tests to detect differences. In general, the multivariate analysis of tumor multiplicity data achieved a greater level of significance than any of the univariate response measures and procedures.

The methods using the univariate response measures for tumor count data have limitations which could possibly bias the analysis and lead to erroneous results. These methods assume that an animal's tumor observation is censored at its value at the time of death. Thus, animals dying at different times with equal responses are treated equivalently. Intuitively, tumor free deaths at the end of the study should be greater evidence for no treatment effect than those earlier in the study. Similarly, death early in the study with a specific tumor response should be greater evidence for a toxic effect than an animal dying later with the same response. The multivariate response measure considered is the only measure examined that is not biased by any of the characteristics of the data. It accounts for mortality and time of tumor occurrence. While intuitively this method seems most reasonable, it has several disadvantages: several assumptions are necessary for the validity of the test and computationally it is the most difficult in that the computer software is not readily available.

Each of the univariate response measures contains important information that can be used to interpret the results of a study, but each also has limitations and thus, it is recommended that the analysis of the mouse skin assay data should not be confined to a single response measure. A thorough analysis should include an examination and test for equality of survival distributions, a comparison of time to first tumor distributions and a comparison of tumor multiplicity. The strong performance of the Gehan Wilcoxon multiple comparison test in the analysis of the Sencar tracking system data leads to the recommendation of its usage. As a result of these analyses, if an association between tumor response and dose of a compound is found and it is desirable to examine the data for a monotone dose response, a trend analysis using the Cox-type trend test can be performed. By conducting the analysis in this manner, full use is made of all available information on the toxicity of a compound in a mouse skin painting assay. The recommendations are summarized in Table 3.

Mouse Lung Adenoma Assay. A standardized protocol for the mouse lung adenoma bioassay as a short-term test for the carcinogenic potential of chemicals has been described by Shimkin an Stoner (14,15,16). The recommended protocol includes three dose groups of 20 to 30 mice each in addition to a control group. The dose levels are generally set at the maximum-tolerated dose and at one-half and one-fifth this dose. Strain A mice are preferred because they are the most susceptible strain for this assay. The duration of the experiment is twenty-four weeks for the Strain A mice. All animals are sacrificed at a specified time and the tumors in the lungs counted. Animals that die prior to the final sacrifice are not examined and are omitted from the study. Shimkin and Stoner recommended that the assay include animals of both sexes and several routes of administration because the carcinogenic potential of a compound may depend upon sex and route of exposure.

The lung adenoma response data is less complicated than that from a skin painting assay. The tumor count data are collected at a single point in time and generally the animal survival rates are sufficiently high so that animals dying prior to terminal sacrifice can be omitted from the analysis with negligible effects. The carcinogenic potential of a chemical is tested using two response measures; tumor incidence and tumor multiplicity. Tumor incidence is estimated by the number of animals per treatment group with lung tumors and tumor multiplicity by the number of visible tumors on the lung surface of each animal. Strain A mice have a high incidence of "spontaneous" lung tumors and Stoner and Shimkin (15) have noted that the induction of lung adenomas with chemical carcinogens is directly related to the spontaneous occurrence of lung tumors in the various mouse strains. Thus, the multiplicity of the benign lung tumors, not the tumor incidence rate, is the primary response measure of interest.

Crump et al. (17) studied statistical methods appropriate for the tumor response data collected in the mouse lung adenoma assay. Omnibus tests for differences of any type, tests for a positive response, tests for a monotone dose-response trend, and tests for differences between sexes or between two routes of exposure were evaluated for both tumor incidence data and tumor multiplicity data. Tests based on parametric models as well as distribution-free tests were investigated. Parametric models included the one-hit, negative binomial and logit models. Distribution-free tests included Fisher's and extensions thereof, Mantel-Haenszel, Jonckheere, logrank and Wilcoxon.

The results of the study indicated that no single statistical test will have the greatest power against all possible alternatives. Tests utilizing tumor multi-

plicity data had greater power than tests utilizing tumor incidence data. No advantage to using parametric tests was found and it was recommended that nonparametric tests be used exclusively. Logrank tests, as a class, appeared to perform well whether for omnibus, trend or positive response testing of mouse lung adenoma data. Among the trend test considered, the Mantel-Haenszel test generally had the highest power when there was a monotone dose-response trend, but performed poorly when the effect was not monotone in dose. Thus, the logrank test and, in the case of a monotone dose response, the Mantel-Haenszel test were recommended for testing for differences among exposure groups, between sexes and between routes of administration. The recommendations are summarized in Table 3.

Rat Liver Foci Assay. The liver foci bioassay has been proposed as a short-term assay to test chemicals or mixtures for their ability to act as initiators. A description of the bioassay can be found in Pereira (18). Generally, male Sprague-Dawley rats are used and the test substance is administered orally at six, twelve or eighteen hours after a two-thirds partial hepatectomy. The test substance is administered during the initiating stage of the dosing regimen and a known promoter is administered during the promotion stage. The assay is conducted in two phases. The first phase tests for initiating ability at a dose level no higher than one half the the LD_{50}. If a positive response is found then a dose-response study is conducted to determine if the positive response can be quantitatively related to dose. The histochemical localization of enzyme altered foci, which are assumed to be preneoplastic cells, is the response endpoint for the liver foci assay. Approximately twelve weeks after initiation, the animals are sacrificed and two or three liver sections from each animal are scored for the presence of foci.

Two response measures are of interest in testing the initiating ability of compounds: the number of focal intersections with nine or more nuclei per unit area of tissue examined and the number of focal intersections with nine or more nuclei per unit volume of liver. The first measure is used in both phases of testing and the second only in the second phase to test for a dose-response relationship.

The distribution of the liver foci response data is similar to the lung adenoma data in that it is heavily skewed and contains many zero responses. Therefore, based on the results of the study of statistical methods for the other CTM short-term tests, the proposed approach is to investigate the performance of the nonparametric tests recommended for the lung adenoma assay and the Sencar skin painting assay. Specifically, the nonparametric logrank and Wilcoxon tests will be applied to the data from each of the experiments testing the initial CTM compounds. The performance of

the proposed methods will be compared and evaluated for each response measure. The proposed statistical procedures include a test for positive response, a test for a dose-related trend and an omnibus test for any type of difference among treatment groups. The sensitivity of the liver foci bioassay with respect to detecting known initiators will be evaluated empirically in terms of the response measures.

Estimation of Carcinogenic Potency from Long-term Bioassays

The ultimate goal of the CTM project is to predict the carcinogenic potential of unknown chemicals or complex mixtures with respect to their rankings within a known class of chemicals, based only on the results of short-term in vivo bioassays. In this regard it is imperative that the proposed index of carcinogenic potency and the methods for estimating it be flexible in order to be reliable for complex situations involving many sources of variation. There are three main areas to be considered in the characterization of long-term potency: long-term study data sources, study selection criteria and the method of estimating potency. Each will be discussed individually.

Currently two options exist for long-term study data sources for the initial CTM compounds. A CTM data base can be developed and evaluated from the literature or the information in existing data bases used. Each option has advantages and disadvantages. Developing a specialized CTM database is costly and time consuming but has the advantage in that the development can focus on the information precisely needed for the project. Using existing data bases is an efficient method of gathering information, but the data is limited to that already collected and the completeness and reliability of the data obtained can not be controlled. Methods have been proposed for each option.

In developing a long-term study data base from the literature, study selection would be based on the following minimum criteria:

- Species - rats, mice or possibly hamsters
- Exposure
  - --oral or inhalation route
  - --continuously at a constant rate
  - --for a large portion of life span
- Concurrent controls
- Information available on:
  - —administration of dosages

--sex of animal

--number of animals in each treatment group

Kodell et al. (19) have reviewed various methods for estimating potency reported in the literature. Risk assessment of carcinogens is primarily concerned with low level exposure. Risk estimated at low doses can vary substantially depending on the assumed dose-response model and the dose range of the available data. Thus, it was proposed that carcinogenic potency be represented by:

$$\text{Potency} = 0.10/CD_{10}$$

where CD_{10} is the dose level that will result in an extra risk of 10%. Thus, this potency measure is risk divided by dose, and is consistent with the present EPA measure (the slope at low doses). Associated $100(1-\alpha)\%$ lower and upper confidence limits,

$$0.10/UL_{10} \quad \text{and} \quad 0.10/LL_{10}$$

where UL_{10} (LL_{10}) are $100(1-\alpha)\%$ upper (lower) confidence limits on CD_{10}, will be calculated.

Although it was proposed to estimate potency based on the dose level that would result in an extra 10% risk, this level could easily be modified, eg., 25%, and the stability of the resulting relative potency estimates evaluated. Additionally, various α levels of the associated confidence interval could be calculated and compared.

The dominant theory of the carcinogenic process has resulted in the development and use of the multistage model for chemical carcinogen risk assessment. The flexibility of the multistage model enables it to fit dose-response models of various shapes. The multistage model is defined as:

$$P(d) = 1 - \exp(- q_0 - q_1 d - \ldots - q_k d^k)$$

$$q_i \geq 0, \quad i = 0, 1, 2, \ldots, k$$

where P(d) is the probability of a carcinogenic response at dose d of the carcinogen and k, $q_0, \ldots, q_k$ are parameters. In practice, k is set equal to the number of dose levels minus one. This model will be used to estimate the dose level that would

result in the excess risk of interest. For consistency across studies involving a different number of dose groups, it is proposed to fit the multistage model with degree equal to the minimum of three and the number of dose groups minus one. However for confidence intervals, a third degree multistage model would be used regardless of the number of doses, as the solution would always be unique. A chi-square goodness-of-fit test will be performed to evaluate the fit of the model to the data.

An existing literature data base that could be used for the CTM project was published by Gold et al. and Peto et al. in 1984 (20,21). The following criteria were used to select studies from the "Carcinogenic Potency Database" to be used in the initial evaluation of the CTM project:

- Species - mouse or rat
- Exposure
  - --minimum of 15 months dosing
  - --oral route of exposure (food, water or gavage)
- Minimum study duration of 15 months
- Tumor site - all solid tumors

Gold et al. and Peto et al. propose that a numerical measure of carcinogenic potency based on the TD_{50} be used as the standard index and defined the TD_{50} as follows:

> 'For any particular sex, strain, species and set of experimental conditions, the TD_{50} is the dose rate (in mg/kg body weight/day) that, if administered chronically for a standard period - the "standard lifespan" of the species - will halve the mortality-corrected estimate of the probability of remaining tumorless throughout that period.'

The articles contain detailed discussions of the TD_{50}, the procedure for estimating the TD_{50} and its confidence limits, and associated problems.

For expediency and purposes of comparison, the initial evaluation of the ability of the three short-term assays to predict relative carcinogenic potency will be done using the Gold et al. data base and its standardized potency index. Based on the results, the effect level at which potency is measured will be varied (eg. TD_{10}) and the predictability and stability of the relative of potency estimates using the three short-term assays re-evaluated. The proposed methods are summarized in Table 3.

Estimation of Potency from Short-term Bioassays

Basically there are three types of tumor response data in the short-term studies currently being evaluated for the CTM. The first type is quantal data which counts the number of animals with at least one tumor in each dose group. The response data from each of the assays can be summarized in this manner. The second type is tumor multiplicity data, that is the number of tumors observed for each animal. Data from the Sencar mouse skin painting assay and the Strain A mouse lung adenoma assay can be represented by this type of quantitative tumor count. The third type of response data can be described as a "continuous" data measure and is applicable to the rat liver foci assay. For this assay the response for each animal can be measured as the number of tumors (foci) per unit area or volume of liver examined.

Kodell et al. (19) have proposed methods for estimating potency for each type response data. For quantal data, the potency will be defined and estimated by the same methods proposed for long-term studies compiled in a CTM data base. For tumor multiplicity data and "continuous" response data, the mean response is assumed to be linear in dose and the potency is defined as the slope of the straight line. The response is assumed to follow a negative binomial or Poisson distribution. The potency and its associated confidence interval will be estimated by the method of maximum likelihood. A chi-square goodness-of-fit test will be performed to evaluate the fit of the model to the data. The proposed procedure tests for deviations from linearity in dose of the mean response. If a statistically significant deviation is found the model will be fit to the data excluding the highest dose group. If the response for the highest dose group is below that predicted by the new model then it is omitted from the model. The highest dose group is also omitted if its response is above the fitted model and the response for at least one other dose group is significantly increased above the control response. Otherwise the highest dose group is included even though there is a significant lack of fit. If the highest dose group is omitted then the model fitting procedure is repeated on the remaining dose groups until no dose groups can be omitted. The reason for not always excluding the highest dose group when the model is not adequate is that in in doing so the data that give evidence of a carcinogenic effect may also be omitted. The proposed methods are summarized in Table 3.

Prediction of Long-term Potency from Short-term Potency

Methods. Potency prediction can be defined quantitatively based on the point estimates of potency or qualitatively based on the ranks of the potency estimates.

Since an objective of this initial phase of the CTM project is to investigate the ability of the three short-term assays to predict relative long-term potency, it is proposed to characterize the potency of one compound to another based on a generalized ranking scheme. In order to take into account random variation and differential study sizes, both of which could inappropriately influence ranking, a ranking procedure based on both upper and lower confidence limits has been developed (19).

Two procedures for ordering the confidence intervals have been defined. The first, a strong ordering, defines two confidence intervals as ordered if they do not overlap and tied if there is any overlap. The second, a weak ordering, defines two confidence intervals as ordered if the upper and lower limit of one interval are above (below) the upper and lower limits, respectively, of the other and tied if one interval falls completely inside the other. The weak ordering method will result in the ordering of many pairs of confidence intervals which were unorderable using the strong ordering method. Both methods will result in a partially ordered set, the refinement of which can be adjusted by altering the confidence coefficient.

A generalized ranking algorithm has been derived for this phase of the CTM project and will be applied to the partially ordered confidence intervals on the potencies for both the short-term and long-term studies. The algorithm can be defined as follows:

For $i=1,2,\ldots,N$, where N is the number of "points" in the partially ordered set to be ranked, let

n_{io} = number of points that are tied (in a broad sense) with point i, including point i itself,

n_{ig} = number of points that are strictly greater than point i,

n_{is} = number of points that are strictly less than point i.

Then the generalized rank for point i, with lower points being assigned lower ranks, is defined as

$$R_i = [(n_{is} + 1) + (n_{is} + 2) + \ldots + (n_{is} + n_{io})]/n_{io}$$

$$= n_{is} + (n_{io} + 1)/2.$$

R_i is simply the average of the ranks that would have been assigned to the points that are tied with respect to the point being ranked, if there had been no ties and if every point greater (smaller) than the point being ranked was also greater (smaller) than those points.

Relative potency estimates based on confidence intervals have advantages over those based on point estimates or their ranks when the ranges of true potencies are very narrow, when some chemicals are clustered together, and when the estimates are based on poor data. In such situations the confidence intervals will be substantially overlapped. Correlation analysis using point estimates or ranks would mask such situations and likely be statistically non-significant.

Some additional issues that need to be considered include the selection of the carcinogenic response, i.e., tumor type and site, from the long-term studies to be used to estimate potency, the combination of information from several long-term studies for a single potency estimate, and the combination of information from the three short-term assays for a single potency estimate.

At this point in time, for the long-term studies, it is proposed to use the tumor site from a single study that gives the lowest TD_{50} or equivalently the highest upper limit on potency for a given compound. Thus the point estimate of potency as well as its confidence limits will be based on a single study's results for one species, strain and sex combination. Characterization of a chemical's long-term carcinogenic potency will be based on this value. If it becomes desirable to combine information from several studies, the individual point estimates and confidence intervals of potencies could be combined using a weighted average on a log scale as follows:

$$A = \prod_i a_i^{\omega_i}$$

where a_i - potency point estimate or confidence limit from study i

ω_i - $n_i / \sum_j n_j$

n_i - total sample size for study i.

As defined above the weights are based on total sample size, but weights based on other significant study factors would be investigated. Prior to combining, the

data would be adjusted to standardize for other differences in protocol, such as duration of exposure and duration of study.

Initially a separate estimate of potency will be made for each route and sex of the three short-term assays. These will be compared individually to the potency estimates from the long-term studies. It is expected that each of the short-term assays potentially will be selective in its response to certain chemicals or classes of chemicals. If it is desirable to combine information from several of the short-term tests to predict long-term potency, the use of weighted multiple regression methods to combine the generalized ranks will be examined.

Evaluation. In evaluating the ability of the short-term assay to predict long-term potency it should be kept in mind that it is the overall correlation from the CTM that is of main interest. Although relative ranking is the main concern in this phase of the project, point estimates of potency will also be evaluated for comparative purposes. Three different techniques will be used in the evaluation. The first step will be a graphical evaluation; plotting short-term vs long-term potency estimates and their confidence intervals. This is a descriptive evaluation to aid in ensuring that obvious relationships are not overlooked. Formal testing for a correlation between the generalized ranks obtained from the confidence intervals of potencies will be done using Pearson's Product Moment Correlation Coefficient and Kendall's Tau. Pearson's correlation coefficient measures the strength of the linear relationship between the generalized ranks of the long-term and short-term potencies and Kendall's Tau measures the association in terms of concordance of general order regardless of the absolute value of the ranks. The point estimates of potency can be compared using these two correlation measures as well as Spearman's Rho, a rank correlation. Further evaluation of the CTM can be done by isolating each chemical one at a time as an unknown and predicting its relative rank among the other chemicals with respect to long-term carcinogenic potency. These results can be compared to the actual rankings based on the data from the initial CTM compounds. In this way the adverse influence of factors unrelated to the basic methodology can be assessed.

Various problems inherent in this type of project will need to be evaluated and addressed as work progresses. Of necessity, this will be an evolutionary process. Based on the results of the initial phase, the proposed methodology can be modified as necessary and the strategy for future work developed.

The research described in this article (paper/chapter, report) has been reviewed by the Health Effects Research Laboratory, United States Environmental Protection Agency and approved for publication. Approval does not signify that the contents necessarily reflect the views and policies of the Agency nor does the mention of trade names or commercial products constitute endorsement or recommendation for use.

Table 3

Summary of Statistical Methods for CTM Project

Short-term Assay	Type Data	Statistical Test Positive	Statistical Test Trend	Potency Estimation Method
Sencar Skin Painting	Time to 1st tumor	Gehan Wilcoxon	Cox Type Trend	NA
	Tumor Multiplicity			
	Univariate	Gehan Wilcoxon	Cox Type Trend	Linear regression on mean response (slope)
	Multivariate	---	Cox Type Trend	NA
Strain A Lung Adenoma	Tumor Multiplicity	Logrank	Mantel-Haenszel Trend Test	Linear regression on mean response (slope)
Rat Liver Foci	"Continuous" (Foci/sq cm or Foci/cu cm)	Logrank* Gehan-Wilcoxon*		Linear regression on mean response (slope)

Long-term Assays	Potency Estimation Method
Gold et al. Database	Based on Standardized Index (TD_{55}**)
CTM Literature Database	Based on CD_{xx}+ (Multistage Model)

* Statistical Methods currently being evaluated.

** TD_{50} - Chronic dose rate resulting in reducing the probability of remaining tumor free during a lifetime by 50%

+ CD_{xx} - Dose rate which results in an excess risk of developing a tumor of xx% over the concurrent controls.

REFERENCES

1. R. J. Bull and M. A. Pereira. "Development of a Short-term Testing Matrix for Estimating Relative Carcinogenic Risk." Journal of the American College of Toxicology, Vol. 1, No. 1, 1982, pp. 1-15.

2. Journal of the American College of Toxicology, Mary Ann Liebert, Inc., New York, Vol. 1 No. 1, 1982.

3. T. J. Slaga, S. M. Fischer and L. L. Triplett. "Comparison of Complete Carcinogenesis and Tumor Initiation and Promotion in Mouse Skin: The Induction of Papillomas by Tumor Initiation-Promotion a Reliable Short Term Assay. Journal of the American College of Toxicology, Vol. 1, 1982, pp. 83-99.

4. R. J. Bull, M. Robinson, R. D. Laurie, et al. "Carcinogenic Effect of Acrylamide in Sencar and A/J Mice." Cancer Research, Vol. 44, 1984, pp. 107-111.

5. K. S. Crump and T. Ng. "A Study of Statistical Methods for Testing for a Dose-Related Effect in Skin Painting Studies." Response report for for U.S. EPA, Health Effects Research Laboratory, Cincinnati, Ohio, Meta Systems, Inc., 1983.

6. E. A. Gehan. "A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples." Biometrika, Vol. 52, 1965, pp. 203-222.

7. N. Mantel. "Ranking Procedures for Arbitrarily Restricted Observations." Biometrics, Vol. 23, 1967, pp. 65-78.

8. R. E. Tarone and J. Ware. "On Distribution-free Tests to Equality of Survival Distributions." Biometrika, Vol. 64, 1977, pp. 156-160.

9. N. Mantel. "Chi-square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure." Journal of American Statistical Association, Vol. 58, 1963, pp. 690-700.

10. D.R. Cox. "Regression Models and Life Tables." Journal of the Royal Statistical Society, Series B, Vol. 34, 1972, pp. 187-220.

11. R. E. Tarone. "Test for Trend in Life Table Analysis." Biometrika, Vol. 62, 1975, pp. 679-682.

12. R. L. Prentice, B. J. Williams and A. V. Peterson. "On the Regression Analysis of Multivariate Failure Time Data." Biometrika, Vol. 68, 1981, pp. 373-379.

13. J. A. Stober. "Methodology for Statistical Analysis of Sencar Mouse Skin Assay Data." Environmental Health Perspectives, in press.

14. M. Shimkin and G. Stoner. "Lung Tumors in Mice: Application to Carcinogenesis Bioassay." Advanced Cancer Research, Vol. 21, 1975, pp. 1-58.

15. G. D. Stoner and M. B. Shimkin. "Strain A Mouse Lung Tumor Bioassay." Journal of the American College of Toxicology, Vol. 1, No. 1, 1982, pp.145-169.

16. G. D. Stoner and M. B. Shimkin. "Lung Tumors in Strain A Mice as a Bioassay for Carcinogenicity." Handbook of Carcinogenesis Testing, H. Milman and E. K. Weisburger, (eds). Park Ridge, New Jersey, Noyes Publications (in press), 1985.

17. K. S. Crump, T. Ng and R. L. Kodell. "Statistical Methods for Analyzing Mouse Lung Adenoma Bioassay Data." Response Report for U.S. EPA, Health Effects Research Laboratory, Cincinnati, Ohio, Washington Consulting Group, 1985.

18. M. A. Pereira. "Rat Liver Foci Bioassay." Journal of the American College of Toxicology, Vol. 1, No. 1, 1982, pp.101-117.

19. R. L. Kodell, T. Ng and K. S. Crump. "Statistical Methods for Predicting Long-term Carcinogenic Potencies for Organic Chemical Compounds from Short-term in vivo Tests." Response report for U.S. EPA Health Effects Research Laboratory, Cincinnati, Ohio, K. S. Crump and Company, Inc., May 1985.

20. L. S. Gold, C. B. Sawyer, R. Magaw et al. "A Carcinogenic Potency Database of the Standardized Results of Animal Bioassays." Environmental Health Perspectives, Vol. 58, 1984, pp. 9-319.

21. R. Peto, M. C. Pike, L. Bernstein et al. "The TD_{50}: A proposed General Convention for the Numerical Description of the Carcinogenic Potency of Chemicals in Chronic-Exposure Animal Experiments". Environmental Health Perspectives, Vol. 58, 1984, pp. 1-8.